Subgroup Analysis: Principles and Applications

Andrea Z. LaCroix, PhD
Fred Hutchinson Cancer Research Center
Professor of Epidemiology
University of Washington

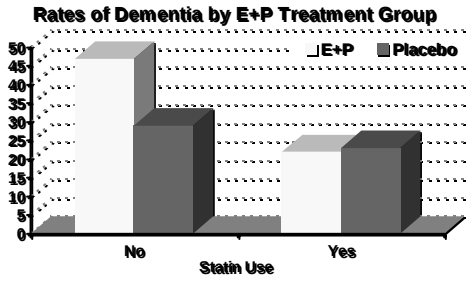
Purpose of Subgroup Analysis

To determine if there are subgroups of trial participants, for which the treatment is more (or less) effective (or harmful) than is indicated by the overall comparison.

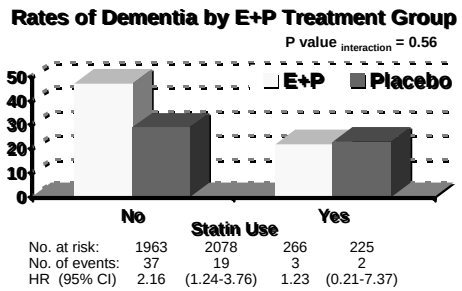
"We have a scientific and ethical obligation to try and identify such subgroups."

Source: Pocock SJ, Assmann SE, Enos LE, Kasten LE. Stat Med 2002;21:2917-30.

Is this an important subgroup difference?



Is this an important subgroup difference?



No. at risk:	1963	2078	266	225
No. of events:	37	19	3	2
HR (95% CI)	2.16	(1.24-3.76)	1.23	(0.21-7.37)

Common Objectives of Subgroup Analysis

- Provide supporting evidence for main findings
- Check on the consistency of the main findings
- Address particular concerns or evaluate the efficacy and safety of a treatment in a specific patient subgroup
- Generate hypotheses for future studies

Source: Cui L, Hung H, Wang S, Tsong Y. J Biopharm Stat 2002;12:347-358.

Inappropriate Uses of Subgroup Analysis

1. Rescue of a negative trial: Mission is to find some group who benefits.
2. Rescue of a harmful trial: Mission is to find some group who is not harmed (or better yet, benefits).
3. Data dredging: Mission is to find “interesting” results without a pre-specified analysis plan or hypotheses.

How to avoid inappropriate uses of subgroup analysis

1. Develop a pre-specified analysis plan that states how subgroup analyses will be conducted.
2. Specify in advance any hypotheses to be tested about treatment effects in subgroups based on prior evidence and plan for adequate power in these subgroups.
3. Avoid conducting subgroup analyses with any of the aforementioned “missions.”

Definitions

Proper Subgroup: A group of patients characterized by common set of “baseline” characteristics.

Improper Subgroup: A group of patients characterized by a variable measured *after* randomization *and* potentially affected by treatment.

Source: Yusuf S, Wittes J, Probstfield J, Tyroler H. JAMA 1991; 266:93-98.

Definitions

Interaction: Treatment effects that differ by subgroup.

Quantitative interaction: When a treatment effect is beneficial or harmful in all subgroups, but the magnitude of effect varies among subgroups.

Qualitative interaction: When the treatment effect is truly beneficial in some subgroups, but truly harmful in others.

Source: Yusuf S, Wittes J, Probstfield J, Tyroler H. JAMA 1991; 266:93-98.

Subgroup Analysis: Five big problems

1. Statistical power is limited to detect differences in response to treatment in subgroups (Type II errors).
2. If many subgroup analyses are performed (with or without pre-specification) the possibility of finding "interesting" (significant) results by chance alone is large (Type I errors).
3. Appropriate statistical tests for making inferences from subgroup analyses are often not used.

Subgroup Analysis: Five big problems

4. Within a subgroup, treatment group comparability can be compromised, creating an imbalance in prognostic factors and selection bias.
5. How much subgroup analyses should affect the interpretation and conclusions of the overall trial results is debatable.

Source: Pocock SJ, Assmann SE, Enos LE, Kasten LE. Stat Med 2002; 21:2917-30.

"…any particular subgroup finding, no matter how intriguing, is prone to be an exaggeration of the truth."

Source: Pocock, Assmann, Enos, Kasten. Stat Med 2002; 21:2917-30.

Big Problem #1: Low Power

1. A well designed trial is large enough to detect a clinically significant overall difference.
2. Unless the trial is specifically designed to have sufficient power within subgroups of interest, it cannot be expected to detect effects within even large subgroups, and is unlikely to detect interactions.

Source: Yusuf S, Wittes J, Probstfield J, Tyroler H. JAMA 1991;266:93-98.

Big Problem #2: Too Many Statistical Tests

1. The more questions asked, the greater the chance of finding a significant result that is not true.
2. Reported p-values often bear little relationship to the true probability of finding a chance effect.

Source: Yusuf S, Wittes J, Probstfield J, Tyroler H. JAMA 1991;266:93-98.

Big Problem #2: Too Many Statistical Tests

3. "Partitioning the dataset into many small subsets will almost ensure the discovery of a suggestive, though not necessarily statistically significant, treatment effect."
4. Proving uniformity of effect across many subgroups is also extremely difficult.

Source: Yusuf S, Wittes J, Probstfield J, Tyroler H. JAMA 1991;266:93-98.

Big Problem #3: Lack of Appropriate Statistical Tests

1. Most useful approach for evaluating subgroup treatment differences is performing statistical tests for interaction.
2. Only 15 of 35 reports used interaction tests.
3. Reporting subgroup p-values and confidence intervals are misleading since some likely will and will not be significant based on the size of the subgroup and chance.

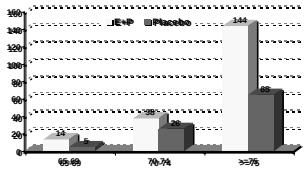
Source: Pocock SJ, Assmann SE, Enos LE, Kasten LE. Stat Med 2002; 21:2917-30.

CHD Outcomes (Annualized Percentages) by Self-Reported History of CHD Related Conditions

	Estrogen + Progestin	Placebo	Hazard Ratio	95% Nominal CI
No prior MI or CABG/PTCA	145 (0.34%)	106 (0.26%)	1.28	(1.00, 1.65)
Prior MI or CABG/PTCA	19 (2.08%)	16 (1.60%)	1.28	(0.64, 2.56)

WOMEN'S HEALTH INITIATIVE

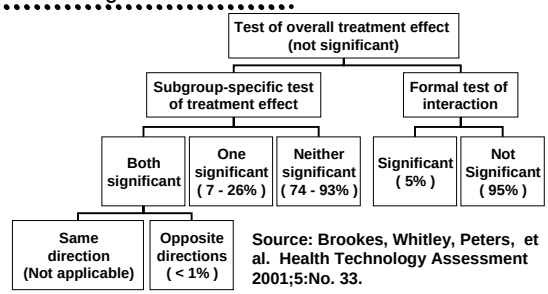
Rates of Dementia per 10,000 person-years According to E+P Treatment Assignment



P-value for interaction = 0.60.

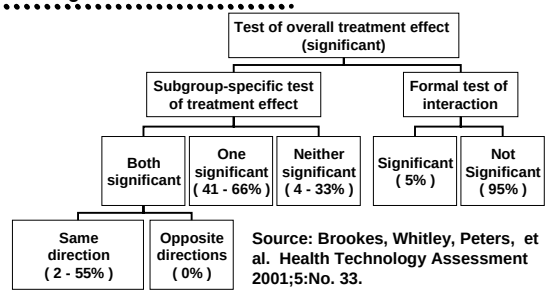
No. of events:			
E+P	6	12	22
Placebo	2	9	10
HR	3.25	1.47	2.34
95% CI	0.66-16.1	0.62-3.49	1.11-4.94

Summary of Simulation Results with No Significant Overall Treatment Effect



Source: Brookes, Whitley, Peters, et al. Health Technology Assessment 2001;5:No. 33.

Summary of Simulation Results with Significant Overall Treatment Effect



Source: Brookes, Whitley, Peters, et al. Health Technology Assessment 2001;5:No. 33.

Big Problem #4: Compromised Treatment Group Comparability

1. Can arise if randomization is not stratified by the factor defining the subgroup.
2. Can arise with stratified randomization if the subgroup is small.
3. More likely to occur for prognostic factors with a prevalence rate approaching 50%.
4. Severity of the imbalance and correlation of the prognostic factor with the outcome are the main determinants of the extent of selection bias.

Source: Cui L, Hung H, Wang S, Tsong Y. J Biopharm Stat 2002;12:347-358.

Big Problem #4: Compromised Treatment Group Comparability

5. Difficult if not impossible to detect (factors may be unmeasured, operate jointly with other factors, give no clear signal).
6. Direction of the bias is generally unknown.
7. Stratified randomization and sufficiently large (>100 patients/treatment group) are the best defense.

Source: Cui L, Hung H, Wang S, Tsong Y. J Biopharm Stat 2002;12:347-358.

Big Problem #5: Interpretations of Subgroup Analyses are Subjective and Contentious

How much emphasis to put on a subgroup finding should be based on:

- Strength of the statistical evidence for interaction +
- Wise judgment
- Over interpretation is common!

Source: Pocock, Assmann, Enos, Kasten. Stat Med 2002;21:2917-30.

What About Biologic Plausibility?

- Another reasonable criteria for informing wise judgment?

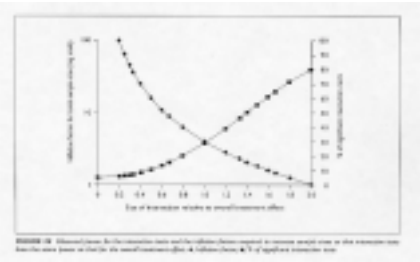
  "The human mind is sufficiently fertile that there is no shortage of biologically plausible explanations or indirect evidence to support almost any observation."

Source: Oxman AD, Guyatt GH. Ann Intern Med 1992;116:78-84

Pre-specified vs. Post hoc Comparisons: Not Such a Big Problem

- A large number of comparisons whether pre-specified in the protocol or planned after the initial analyses (post hoc) are still vulnerable to all of the problems reviewed.
- Only those subgroup analyses that are motivated by hypotheses generated from prior studies and planned for in the design of the trial to ensure adequate statistical power are truly pre-specified.

Power and Inflation Factors for Pre-specified Subgroup Analyses



Source: Brookes, Whitley, Peters, et al.
Health Technology Assessment 2001; 5:No. 33.

Common Problems in Reporting the Results of Subgroup Analyses

1. Presentation of survival curves or incidence rates as main depiction of subgroup differences.
2. Selective presentation of the more interesting subgroup results, selective omission of many others.
3. Failure to conduct and/or report tests of statistical interaction.
4. Reliance on subgroup p-values instead of interaction tests to gauge the statistical significance of the finding.
5. Failure to report, or even count, the number of subgroups examined during analyses.

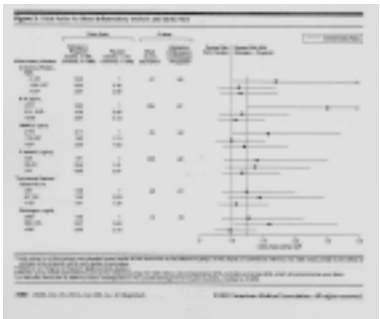
Approach to Subgroup Analysis Used in WHI Priority Papers

- No evidence-based subgroup hypotheses were stated in the protocol.
- Protocol stated that subgroup analyses by major disease risk factors would be explored (e.g., age, BMI)
- Investigators defined subgroups of interest after initial analyses were conducted.

Approach to Subgroup Analysis Used in WHI Priority Papers

- Interaction tests were required.
- Reporting number of subgroups examined and number of significant comparisons expected based on chance alone was also required.

Example: WHI Stroke Paper (Smoller, 2003)



Problems Encountered in WHI Subgroup Analyses

- Counting number of subgroup comparisons was not done systematically, especially for tests that were not reported.
- Some investigators explored second and third level subgroups.
- A single baseline variable could be redefined and tested numerous times.

Problems Encountered in WHI Subgroup Analyses

- Wise judgment was subjective and contentious as promised.
- Journal editorial review could not be relied upon to resolve disagreements.

Our Recommendations

- Consider all subgroup analyses that were not specified in the protocol as post hoc.
- When depicting subgroup analyses, show the reader all of the data needed to interpret the result.
- Do not selectively pick subgroup results that appear interesting and fail to inform the reader about other tests conducted.
- Always reports tests of interaction.
- Always report the number of tests conducted.

Our Recommendations

- Presentation of incidence rates or survival curves can be misleading because these curves are not accompanied by error bars.
- Subgroups with less that 100 subjects per treatment group should be regarded as very prone to error.
- Be zealous in avoiding over interpretation of subgroup results. Think carefully about results that appear in the abstract and statements that suggest translation to medical practice.

The Bottom Line

- Even when the statistical evidence for interaction is very strong, it can be impossible to determine which significant interactions are real and which are due to chance.
- Testing these interactions in other trials is the best way to determine if the result is real.
- Above all, do no harm with subgroup analysis!